基于局部结构学习的非线性属性选择算法 *

李佳烨[†]，张乐园，雷　聪，甘江璋，吕治政

(广西师范大学 广西多源信息挖掘与安全重点实验室，广西 桂林 541004)

摘　要：针对大多数高维数据之间不仅有相似性，而且还有非线性关系等特点，提出一种基于局部结构学习的非线性属性选择算法。该算法首先通过核函数把数据映射到高维空间，在高维空间中表示出数据属性之间的非线性关系；然后在低维空间中通过局部结构学习来充分挖掘属性之间的相似性，同时通过低秩约束来排除噪声的干扰；最后通过稀疏正则化因子来进行属性选择、核函数映射来找出数据属性之间的非线性关系、局部结构学习来找出数据属性之间的相似性。该算法是一种嵌入了局部结构学习的非线性属性选择算法。实验结果表明，该算法相比其他的对比算法，有更好的效果。

关键词：属性选择；核函数；低秩；局部结构学习；稀疏正则化

中图分类号：TP301.6　　doi: 10.19734/j.issn.1001-3695.2018.07.0524

Nonlinear feature selection algorithm via local structure learning

Li Jiaye[†], Zhang Leyuan, Lei Cong, Gan Jiangzhang, Lyu Zhizheng

(*Guangxi Key Lab of Multi-source Information Mining & Security*，*Guangxi Normal University, Guilin Guangxi* 541004, *China*)

Abstract: Due to that most high-dimensional data not only has the similarities, but also nonlinear relationships. This paper proposed a nonlinear feature selection algorithm based on local structure learning. Firstly, the algorithm mapped the data to high-dimensional space through kernel functions, and expresses the nonlinear relationship between data features in high-dimensional space. Then, it exploited the similarity between the features in the low-dimensional space through local structural learning. At the same time, it eliminated the interference of noise by the low-rank constraint. Finally, it selected features by sparse regularization factors. It find the non-linear relationships between data features by the kernel function, and find the similarities between the data attributes as the local structure learning. The algorithm is a nonlinear feature selection algorithm embedded with local structure learning. Experimental results show that the algorithm has better results than other comparison algorithms.

Key words: feature selection; kernel function; low-rank; local structure learning; sparse regularization

0　引言

随着计算机与科学技术的发展，信息时代来临，与此同时带来了大量的高维数据[1]。人工智能，数据挖掘等领域也蓬勃发展。人们面对成千上万的数据处理起来会十分困难，而且有时还会出现“维数灾难”等问题[2, 3]。针对这些高维数据，人们必须对其进行预处理。而属性选择便是其中一种十分有效的方式[4]。通过属性选择来对数据进行预处理从而缩小数据维度是十分必要的[5]。

属性选择[6]包括线性属性选择和非线性属性选择，它们的根本目的都是寻找一个相对较小且具有代表性的属性子集。常用的属性选择方法有很多[7]，但它们都不能挖掘出数据属性之间的非线性关系。局部结构学习刚开始是运用到样本上，通过构建样本之间的相似矩阵来充分体现样本之间的结构[8]，从而达到较好的实验效果。但它不能充分体现属性之间的结构关系。为此，本文通过核函数把数据的每一个属性映射到高维空间，从而使它们之间的非线性关系在高维空间中线性可分；与此同时再把局部结构学习运用到数据属性上，从而更好的表示出数据属性在低维空间中的局部结构关系。提出了一种更加有效的属性选择算法，称作基于局部结构学习的非线性属性选择算法（nonlinear feature selection algorithm via local structure learning，缩写为 LS_NFS）。

本文首先通过核函数对数据处理，得到核矩阵，从而解决了只能进行线性属性选择的限制；其次对数据属性构建相似矩阵来进行局部结构学习，可以提高分类准确率；其中还在模型中进行了低秩约束，低秩约束可以排除噪声的干扰；最后嵌入一个向量的 l_1 – 范数来进行属性选择。由于本文同时考虑了数据属性之间的非线性关系和相似性，所以比单一的线性属性选择方法具有更好的效果。经实验验证，该算法在分类准确率上能达到较好的效果。

收稿日期：2018-07-15；**修回日期：**2018-08-27　**基金项目：**国家自然科学基金资助项目（61170131，61263035，61573270，90718020）；国家重点研发计划资助项目（2016YFB1000905）；国家“973”计划资助项目（2013CB329404）；中国博士后科学基金资助项目（2015M570837）；广西自然科学基金资助项目（2015GXNSFCB139011，2015GXNSFAA139306）

作者简介：李佳烨（1993-），男，山西晋城人，硕士研究生，主要研究方向为数据挖掘、机器学习(jiaye_ligxnu@126.com)；张乐园（1995-），男，安徽蒙城人，硕士研究生，主要研究方向为机器学习、数据挖掘；雷聪（1991-），男，湖北大冶人，硕士，主要研究方向为数据挖掘、机器学习；甘江璋（1994-），男，湖南衡阳人，硕士研究生，主要研究方向为机器学习和数据挖掘；吕治政（1992-），男，湖北黄冈人，硕士研究生，主要研究方向为机器学习和数据挖掘.

1 相关理论背景

1.1 核函数

核函数在很早以前就被引入到了机器学习领域，它的引入在一定程度上减轻了“维数灾难”，大大减小了计算量。只要一个对称函数在任意数据集上产生的核矩阵是半正定的，那么这个函数就是核函数。它的形式和参数是多种多样的，通过选取不同的核函数形式和参数，可以改变从低维空间到高维空间的映射，进而对高维空间的性质产生影响，最终改变各种核函数的性能。

常用的核函数有高斯核函数、多项式核函数、感知器核函数、样条核函数等。由于高斯核函数相比其他三种核函数大多数情况下都具有良好的性能，所以本文算法用高斯核函数。即

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2}{2\sigma^2}) \tag{1}$$

其中：$\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 分别表示第 i 和第 j 个属性。σ 为高斯核的宽度参数，它控制了函数的径向作用范围。

因此，本文通过利用高斯核函数把数据属性映射到核空间，进而充分挖掘数据属性之间的非线性关系。

1.2 局部结构学习简介

前人已经证明通过建立数据之间的局部结构可以起到降维的作用[9]，因此本文通过在低维空间中建立数据属性之间的相似矩阵来进行局部结构学习。

假设给定样本数据集 $\boldsymbol{X}\in\boldsymbol{R}^{n\times d}$，其中 n 和 d 分别表示样本数和属性数。通过局部结构学习得到下面的式子：

$$\min_{\boldsymbol{W}}\sum_{i,j}^{d}\|x_i^T\boldsymbol{W}-x_j^T\boldsymbol{W}\|_2^2 s_{i,j} \tag{2}$$

其中：$x_i\in R^{n\times 1}$ 表示第 i 个属性，$\boldsymbol{W}\in\boldsymbol{R}^{n\times d}$ 是一个高维数据在低维空间中的转换矩阵，$s_{i,j}$ 是矩阵 $\boldsymbol{S}$ 的一个元素，表示属性 x_i 和属性 x_j 的相似性。如果属性 x_i 是属性 x_j 的第 k 个最近的邻居，则 $s_{i,j}$ 的值通过高斯核函数式（1）可得；其他情况 $s_{i,j}$ =0。

2 算法描述和优化

2.1 算法描述

本文首先把数据集 $\boldsymbol{X}\in\boldsymbol{R}^{n\times d}$ 拆成 d 个列向量，每个向量 $x_i\in R^{n\times 1}, i=1,\dots,d$；然后把每个 x_i 中的每个元素看成一个独立的属性值 $x_{ij}\in R, j=1,\dots,n$；并将它们投影到核空间就得到核矩阵 $\boldsymbol{K}^{(i)}\in\boldsymbol{R}^{n\times n}$，即

$$\boldsymbol{K}^{(i)}=\begin{bmatrix} k(x_{i1},x_{i1}) & k(x_{i1},x_{i2}) & \cdots & k(x_{i1},x_{in}) \\ k(x_{i2},x_{i1}) & k(x_{i2},x_{i2}) & \cdots & k(x_{i2},x_{in}) \\ \cdots & \cdots & \cdots & \cdots \\ k(x_{in},x_{i1}) & k(x_{in},x_{i2}) & \cdots & k(x_{in},x_{in}) \end{bmatrix} \tag{3}$$

这样原始的 $\boldsymbol{X}\in\boldsymbol{R}^{n\times d}$ 就变成了 d 个核矩阵 $\boldsymbol{K}^{(i)}, i=1,\dots,d$。

无监督的属性选择算法是为了挖掘数据中更具有代表性的属性。在没有类标签 $\boldsymbol{Y}$ 的情况下，用数据矩阵 $\boldsymbol{X}$ 作为一个响应矩阵，可以更好的保持数据原始特征的内部结构[10, 11]。为了充分挖掘数据属性的非线性关系。得到下面式子：

$$\boldsymbol{X}=\sum_{i=1}^{d}\alpha_i\boldsymbol{K}^{(i)}\boldsymbol{W} \tag{4}$$

其中：$\boldsymbol{W}\in\boldsymbol{R}^{n\times d}$ 表示核系数矩阵；$\boldsymbol{\alpha}\in\boldsymbol{R}^{d\times 1}$ 用来进行属性选择，相当于属性的权重向量；α_i 对应于向量 $\boldsymbol{\alpha}$ 的一个元素；$\boldsymbol{K}^{(i)}\in\boldsymbol{R}^{n\times n}$ 是核矩阵。

为了使 $\boldsymbol{X}$ 取得更好的拟合效果，同时考虑数据属性之间在低维空间上的结构关系，本文得到如下式子：

$$\min_{\boldsymbol{S},\boldsymbol{W},\boldsymbol{\alpha}}\left\|\boldsymbol{X}-\sum_{i=1}^{d}\alpha_i\boldsymbol{K}^{(i)}\boldsymbol{W}\right\|_F^2+\lambda_1\sum_{i,j}^{d}\|x_i^T\boldsymbol{W}-x_j^T\boldsymbol{W}\|_2^2 s_{i,j} \tag{5}$$

由于相似矩阵 $\boldsymbol{S}$ 受参数 σ 的影响变化特别大。为了减少调整参数的次数，同时学习到更有效的相似矩阵，本文把结构学习和低维空间学习相互交替进行，从而达到它们最优的结果。具体得到如下式子：

$$\begin{aligned}&\min_{\boldsymbol{S},\boldsymbol{W},\boldsymbol{\alpha}}\left\|\boldsymbol{X}-\sum_{i=1}^{d}\alpha_i\boldsymbol{K}^{(i)}\boldsymbol{W}\right\|_F^2+\lambda_1\sum_{i,j}^{d}\|x_i^T\boldsymbol{W}-x_j^T\boldsymbol{W}\|_2^2 s_{i,j}+\lambda_2\|s_i\|_2^2\\&s.t., \forall i, s_i^T\boldsymbol{1}=1, s_{i,i}=0, s_{i,j}\ge 0, \text{if } j\in \mathrm{N}(i), otherwise\ 0\end{aligned} \tag{6}$$

其中：λ_1，λ_2 是调优参数，s_i 是相似矩阵 $\boldsymbol{S}$ 的第 i 列，$\|s_i\|_2^2$ 被用来保持旋转不变性。$\boldsymbol{1}$ 代表元素全为 1 的向量，$\mathrm{N}(i)$ 代表第 i 个属性的邻居组成的集合，$s_i^T\boldsymbol{1}=1$ 是为了保持旋转不变性。因此，上式就可以使距离越近的属性对应的 $s_{i,j}$ 值越大，距离越远的属性对应的 $s_{i,j}$ 值越小。

为了排除离群点的干扰，同时去除噪声样本[12]。本文对矩阵 $\boldsymbol{W}$ 加入了低秩约束[13]，即

$$\boldsymbol{W}=\boldsymbol{AB} \tag{7}$$

其中 $\boldsymbol{A}\in\boldsymbol{R}^{n\times r}$，$\boldsymbol{B}\in\boldsymbol{R}^{r\times d}$，$r\le\min(n,d)$，同时本文对矩阵 $\boldsymbol{A}$ 进行正交限制，去充分考虑输出变量之间的相关性，再加入一个 $\boldsymbol{\alpha}$ 的 l_1-范数来进行稀疏学习和属性选择。最后得到最终的目标函数如下：

$$\begin{aligned}&\min_{\boldsymbol{S},\boldsymbol{A},\boldsymbol{B},\boldsymbol{\alpha}}\left\|\boldsymbol{X}-\sum_{i=1}^{d}\alpha_i\boldsymbol{K}^{(i)}\boldsymbol{AB}\right\|_F^2+\lambda_1\sum_{i,j}^{d}\|x_i^T\boldsymbol{AB}-x_j^T\boldsymbol{AB}\|_2^2 s_{i,j}\\&+\lambda_2\|s_i\|_2^2+\lambda_3\|\boldsymbol{\alpha}\|_1\\&s.t., \forall i, s_i^T\boldsymbol{1}=1, s_{i,i}=0,\\&s_{i,j}\ge 0, \text{if } j\in \mathrm{N}(i), otherwise\ 0, \boldsymbol{A}^T\boldsymbol{A}=\boldsymbol{I}\end{aligned} \tag{8}$$

其中：$\boldsymbol{A}^T\boldsymbol{A}=\boldsymbol{I}\in\boldsymbol{R}^{r\times r}$，$\lambda_1$、$\lambda_2$、$\lambda_3$ 是调优参数。核矩阵 $\boldsymbol{K}$ 是通过高斯核函数计算出来，主要作用是把数据映射到核空间，从而挖掘数据属性之间的非线性关系。最后一项 $\boldsymbol{\alpha}$ 的 l_1-范数是对 $\boldsymbol{\alpha}$ 进行稀疏，从而来进行属性选择。$\boldsymbol{\alpha}$ 向量对应的元素的值为零，则表示该属性不选。

本文提出的算法具有以下优点：

a）由于一般的属性选择算法只能寻找出数据属性之间的线性关系，并不能发现数据属性之间的非线性关系。因此本算法把数据矩阵的每一个属性通过核函数分别映射为一个核矩阵，从而在核空间中充分挖掘数据属性之间的复杂非线性关系。进而对数据属性之间的关系挖掘的更加彻底。

b）不同于普通的局部结构学习，只是针对样本计算出一个样本之间相似关系的最优结果。而本算法是针对数据属性，同时把相似矩阵学习和低维空间学习相互交替进行，从而达到最优的属性选择效果。

c）低秩约束可以明显的减少计算量，同时低秩表征着数据的冗余程度。噪声样本会使系数矩阵的秩增加，使用低秩约束可以降低噪声的干扰，同时低秩是对数据全局结构的考虑。从而提高算法的运行效率和分类准确率。

LS_NFS 算法（算法 1）伪代码如下。

输入：训练样本 $\boldsymbol{X}\in\boldsymbol{R}^{n\times d}$，控制参数 $\lambda_1,\lambda_2,\lambda_3$；

输出：分类准确率；

a) 通过训练样本得出类指示矩阵；

b) 通过式(1)和(3)建立每个数据属性对应的核矩阵 $\boldsymbol{K}^{(i)}$；

c) 通过式(6)建立数据属性之间的结构相似矩阵 $\boldsymbol{S}$；

d) 依据所选择的模型调用算法 3 求解全局最优解，得到属性选择向量 $\boldsymbol{\alpha}$；

e) 利用最优解 $\boldsymbol{\alpha}^*$ 对原始属性集 $\boldsymbol{X}$ 进行属性选择后得到的属性集作为样本新的属性集；

f) 对新的属性集构成的样本采用 SVM 分类。

2.2 算法优化

由于目标函数不是共凸的，所以无法直接得到闭式解。因此本文提出一种交替迭代优化方法来求解该问题，具体分以下四步：

1）固定 $\boldsymbol{S}$，$\boldsymbol{B}$，$\boldsymbol{\alpha}$，优化 $\boldsymbol{A}$：

当固定 $\boldsymbol{S}$，$\boldsymbol{B}$，$\boldsymbol{\alpha}$ 之后，优化问题式(8)将变为

$$\min_{A}\left\|\boldsymbol{X}-\sum_{i=1}^{d}\alpha_i\boldsymbol{K}^{(i)}\boldsymbol{AB}\right\|_F^2+\lambda_1\sum_{i,j}^{d}\left\|x_i^T\boldsymbol{AB}-x_j^T\boldsymbol{AB}\right\|_2^2 s_{i,j} \quad s.t.,\boldsymbol{A}^T\boldsymbol{A}=\boldsymbol{I} \tag{9}$$

令 $\boldsymbol{P}=\sum_{i=1}^{d}\alpha_i\boldsymbol{K}^{(i)}$，则式(9)可以写成如下式子：

$$\min_{A}\|\boldsymbol{X}-\boldsymbol{PAB}\|_F^2+\lambda_1\sum_{i,j}^{d}\left\|x_i^T\boldsymbol{AB}-x_j^T\boldsymbol{AB}\right\|_2^2 s_{i,j} \quad s.t.,\boldsymbol{A}^T\boldsymbol{A}=\boldsymbol{I} \tag{10}$$

对式(10)进行化简可得

$$\begin{aligned}&\min_{A} tr(\boldsymbol{X}^T\boldsymbol{X}-\boldsymbol{X}^T\boldsymbol{PAB}-\boldsymbol{B}^T\boldsymbol{A}^T\boldsymbol{P}^T\boldsymbol{X}+\boldsymbol{B}^T\boldsymbol{A}^T\boldsymbol{P}^T\boldsymbol{PAB})\\&+\lambda_1 tr(\boldsymbol{B}^T\boldsymbol{A}^T\boldsymbol{XLX}^T\boldsymbol{AB}), s.t.,\boldsymbol{A}^T\boldsymbol{A}=\boldsymbol{I}\end{aligned} \tag{11}$$

其中 $\mathrm{tr}(\cdot)$ 表示矩阵的迹，$\boldsymbol{L}=\boldsymbol{Q}-\boldsymbol{S}\in\boldsymbol{R}^{d\times d}$ 是一个拉普拉斯矩阵，$\boldsymbol{Q}$ 是一个对角矩阵，它每一列的元素为 $q_{i,i}=\sum_{j=1}^{d}s_{i,j}$。对 $\boldsymbol{A}$ 求导可得：

$$2\lambda_1\boldsymbol{XLX}^T\boldsymbol{ABB}^T-2\boldsymbol{P}^T\boldsymbol{XB}^T+2\boldsymbol{P}^T\boldsymbol{PABB}^T \tag{12}$$

由于对 $\boldsymbol{A}$ 进行了正交限制，可以用参考文献[14]中的方法去优化它。

2）固定 $\boldsymbol{A}$，$\boldsymbol{S}$，$\boldsymbol{\alpha}$，优化 $\boldsymbol{B}$

当固定 $\boldsymbol{A}$，$\boldsymbol{S}$，$\boldsymbol{\alpha}$ 之后，优化问题式(8)将变为

$$\min_{B}\left\|\boldsymbol{X}-\sum_{i=1}^{d}\alpha_i\boldsymbol{K}^{(i)}\boldsymbol{AB}\right\|_F^2+\lambda_1\sum_{i,j}^{d}\left\|x_i^T\boldsymbol{AB}-x_j^T\boldsymbol{AB}\right\|_2^2 s_{i,j} \tag{13}$$

易得式(13)等价于以下式子：

$$\begin{aligned}&\min_{B} tr(\boldsymbol{X}^T\boldsymbol{X}-\boldsymbol{X}^T\boldsymbol{PAB}-\boldsymbol{B}^T\boldsymbol{A}^T\boldsymbol{P}^T\boldsymbol{X}+\boldsymbol{B}^T\boldsymbol{A}^T\boldsymbol{P}^T\boldsymbol{PAB})\\&+\lambda_1 tr(\boldsymbol{B}^T\boldsymbol{A}^T\boldsymbol{XLX}^T\boldsymbol{AB})\end{aligned} \tag{14}$$

对 $\boldsymbol{B}$ 求导，并令其导数为零，则可得

$$\boldsymbol{B}=(\boldsymbol{A}^T\boldsymbol{P}^T\boldsymbol{PA}+\lambda_1\boldsymbol{A}^T\boldsymbol{XLX}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{P}^T\boldsymbol{X} \tag{15}$$

3)固定 $\boldsymbol{A}$，$\boldsymbol{B}$，$\boldsymbol{\alpha}$，优化 $\boldsymbol{S}$

固定 $\boldsymbol{A}$，$\boldsymbol{B}$，$\boldsymbol{\alpha}$ 之后，优化问题式(8)将变为

$$\begin{aligned}&\min_{S}\lambda_1\sum_{i,j}^{d}\left\|x_i^T\boldsymbol{AB}-x_j^T\boldsymbol{AB}\right\|_2^2 s_{i,j}+\lambda_2\|s_i\|_2^2\\&s.t.,\forall i, s_i^T\boldsymbol{1}=1, s_{i,i}=0,\\&s_{i,j}\geq 0, \text{if } j\in \mathrm{N}(i), otherwise\ 0\end{aligned} \tag{16}$$

本文先计算出每两个数据属性之间的欧氏距离来构建所有属性的近邻。如果第 j 个属性不属于第 i 个属性的近邻，则 $s_{i,j}$ 的值为零；否则，通过式(19)求解 $s_{i,j}$ 的值。

与此同时，优化 $\boldsymbol{S}$ 等价于单独地优化每一个 $s_i(i=1,...,d)$，因此本文进一步把优化问题转换为如下式子：

$$\min_{s_i^T\boldsymbol{1}=1,s_{i,i}=0,s_{i,j}\geq 0}\sum_{i,j}^{d}(\lambda_1\left\|x_i^T\boldsymbol{AB}-x_j^T\boldsymbol{AB}\right\|_2^2 s_{i,j}+\lambda_2 s_{i,j}^2) \tag{17}$$

这里令 $\boldsymbol{Z}\in\boldsymbol{R}^{d\times d}$，其中 $\boldsymbol{Z}_{i,j}=\lambda_1\left\|x_i^T\boldsymbol{AB}-x_j^T\boldsymbol{AB}\right\|_2^2$，这样(17)式进一步变成：

$$\min_{s_i^T\boldsymbol{1}=1,s_{i,j}=0,s_{i,j}\geq 0}\left\|s_i+\frac{\lambda_1}{2\lambda_2}\boldsymbol{Z}_i\right\|_2^2 \tag{18}$$

在 KKT 条件下，能够得到如下：

$$s_{i,j}=(-\frac{\lambda_1}{2\lambda_2}Z_{i,j}+\tau)_+ \tag{19}$$

由于每个数据属性都有近邻，这里对每个 $Z_i(i=1,...d)$ 进行降序排列，即 $\hat{Z}_i=\{\hat{Z}_{i,1},...,\hat{Z}_{i,d}\}$，则可知 $s_{i,k+1}=0$，$s_{i,k}>0$。可得

$$-\frac{\lambda_1}{2\lambda_2}\hat{Z}_{i,k+1}+\tau\leq 0 \tag{20}$$

在条件 $s_i^T\boldsymbol{1}=1$ 的限制下能够得到：

$$\sum_{j=1}^{k}(\frac{\lambda_1}{2\lambda_2}\hat{Z}_{i,k}+\tau)=1\Rightarrow\tau=\frac{1}{k}+\frac{\lambda_1}{2k\lambda_2}\sum_{j=1}^{k}\hat{Z}_{i,k} \tag{21}$$

4)固定 $\boldsymbol{A}$，$\boldsymbol{B}$，$\boldsymbol{S}$，优化 $\boldsymbol{\alpha}$

当固定 $\boldsymbol{A}$，$\boldsymbol{B}$，$\boldsymbol{S}$ 之后，优化问题式(8)变成：

$$\min_{\alpha}\left\|\boldsymbol{X}-\sum_{i=1}^{d}\alpha_i\boldsymbol{K}^{(i)}\boldsymbol{AB}\right\|_F^2+\lambda_3\|\boldsymbol{\alpha}\|_1 \tag{22}$$

令：
$$\begin{aligned}&f(\boldsymbol{\alpha})=\left\|\boldsymbol{X}-\sum_{i=1}^{d}\alpha_i\boldsymbol{K}^{(i)}\boldsymbol{AB}\right\|_F^2\\&F(\boldsymbol{\alpha})=f(\boldsymbol{\alpha})+\lambda_3\|\boldsymbol{\alpha}\|_1\end{aligned} \tag{23}$$

注意到 $\|\boldsymbol{\alpha}\|_1$ 是凸但非平滑的。所以使用近似梯度法优化 $\boldsymbol{\alpha}$，本文可以通过下面的规则进行更新迭代 $\boldsymbol{\alpha}$。

$$\boldsymbol{\alpha}_{t+1}=\arg\min_{\alpha}G_{\eta t}(\boldsymbol{\alpha},\boldsymbol{\alpha}_t) \tag{24}$$

$$G_{\eta t}(\boldsymbol{\alpha},\boldsymbol{\alpha}_t)=f(\boldsymbol{\alpha}_t)+<\nabla f(\boldsymbol{\alpha}_t),\boldsymbol{\alpha}-\boldsymbol{\alpha}_t>+\frac{\eta_t}{2}\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_t\|^2+\lambda_3\|\boldsymbol{\alpha}\|_1 \tag{25}$$

在上式中，$\nabla f(\boldsymbol{\alpha}_t)=2\boldsymbol{\alpha}_t^T\sum_{i=1}^{n}(\boldsymbol{M}^{(i)}(\boldsymbol{M}^{(i)})^T)-2\sum_{i=1}^{n}\boldsymbol{X}_i(\boldsymbol{M}^{(i)})^T$，这里的 $\boldsymbol{M}$ 矩阵是在求导过程中产生的。η_t 是一个调优参数，$\boldsymbol{\alpha}_t$ 是第 t 次迭代中 $\boldsymbol{\alpha}$ 的值。

通过忽略式(25)中的独立的 $\boldsymbol{\alpha}$ 可以得到

$$\boldsymbol{\alpha}_{t+1}=\pi_{\eta_t}(\boldsymbol{\alpha}_t)=\arg\min_{\alpha}\frac{1}{2}\|\boldsymbol{\alpha}-\boldsymbol{U}_t\|_F^2+\frac{\lambda_3}{\eta_t}\|\boldsymbol{\alpha}\|_1 \tag{26}$$

其中：$\boldsymbol{U}_t=\boldsymbol{\alpha}_t-\frac{1}{\eta_t}\nabla f(\boldsymbol{\alpha}_t)$，$\pi_{\eta_t}(\boldsymbol{\alpha}_t)$ 是 $\boldsymbol{\alpha}_t$ 在凸集 η_t 上的欧几里德投影，因为 $\|\boldsymbol{\alpha}\|_1$ 具有可分离形式，所以式(26)可以写成如下形式：

$$\alpha_{t+1}^i=\arg\min_{\alpha^i}\frac{1}{2}\|\alpha^i-U_t^i\|_2^2+\frac{\lambda_3}{\eta_t}|\alpha^i| \tag{27}$$

其中：α^i 和 α_{t+1}^i 分别是 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\alpha}_{t+1}$ 的第 i 个元素，则根据式(27)，α_{t+1}^i 可以得到以下闭式解：

$$\alpha^{i*}=\begin{cases}u_t^i-\frac{\lambda_3}{\eta_t}\times sign(u_t^i), & \text{if } \|u_t^i\|>\frac{\lambda_3}{\eta_t}\\ 0, & \text{otherwise.}\end{cases} \tag{28}$$

为了加速式(24)中的近似梯度算法，本文加入了辅助变量：

$$\boldsymbol{V}_{t+1}=\boldsymbol{\alpha}_t+\frac{\beta_t-1}{\beta_{t+1}}(\boldsymbol{\alpha}_{t+1}-\boldsymbol{\alpha}_t) \tag{29}$$

其中 $\beta_{t+1}=\frac{1+\sqrt{1+4\beta_t^2}}{2}$，综上所述，得到算法 2。

算法 2. 优化求解式(22)的伪代码

输入：η_0，$\beta_1=1$，γ

输出：$\boldsymbol{\alpha}$

1. 初始化 t=0，$\boldsymbol{\alpha}_0$ 为一个随机的向量
2. do{
3. while $F(\boldsymbol{\alpha}_t)>G_{\eta t-1}(\pi_{\eta t-1}(\boldsymbol{\alpha}_t),\boldsymbol{\alpha}_t)$,
4. 令 $\eta_{t-1}=\gamma\eta_{t-1}$
5. end while

6. 令 $\eta_t = \gamma\eta_{t-1}$

7. 计算 $\boldsymbol{\alpha}_{t+1} = \arg\min_{\boldsymbol{\alpha}} G_{\eta t}(\boldsymbol{\alpha}, \boldsymbol{V}_t)$

8. 计算 $\beta_{t+1} = \frac{1+\sqrt{1+4\beta_t^2}}{2}$

9. 计算式(29) }

10. while (convergence)

算法 3 优化求解式(8)的伪代码

输入：训练样本 $\boldsymbol{X} \in \boldsymbol{R}^{n\times d}$，控制参数 $\lambda_1, \lambda_2, \lambda_3, r$；

输出：$\boldsymbol{A}$，$\boldsymbol{B}$，$\boldsymbol{S}$，$\boldsymbol{\alpha}$；

1. 初始化 t=0；

2. 随机初始化矩阵 $\boldsymbol{A}^{(0)}$ 和 $\boldsymbol{B}^{(0)}$，初始化 $\boldsymbol{S}^{(0)}$ 为全零矩阵；

3. do{

3.1 通过式(9)计算 $\boldsymbol{A}^{(t+1)}$；

3.2 通过式(13)计算 $\boldsymbol{B}^{(t+1)}$；

3.3 通过式(16)计算 $\boldsymbol{S}^{(t+1)}$；

3.4 通过算法 2 计算出 $\boldsymbol{\alpha}^{(t+1)}$；

3.5 更新 $t = t+1$；}

4. while (式(8)收敛)

2.3 算法收敛性证明

根据式(19)，对于所有的 $i, j = 1,\ldots,n$， $s_{i,j}^{(t+1)}$ 有一个闭式解。可以得到以下式子：

$$\begin{aligned}&\left\|\boldsymbol{X} - \sum_{i=1}^{d}\alpha_i \boldsymbol{K}^{(i)}\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\right\|_F^2 + \lambda_1 \sum_{i,j}^{d}\left\|x_i^T\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)} - x_j^T\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\right\|_2^2 s_{i,j}^{(t+1)} + \lambda_2 \sum_{i,j}^{d}\left\|s_i^{(t+1)}\right\|_2^2 \\ &\le \left\|\boldsymbol{X} - \sum_{i=1}^{d}\alpha_i \boldsymbol{K}^{(i)}\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\right\|_F^2 + \lambda_1 \sum_{i,j}^{d}\left\|x_i^T\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)} - x_j^T\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\right\|_2^2 s_{i,j}^{(t)} + \lambda_2 \sum_{i,j}^{d}\left\|s_i^{(t)}\right\|_2^2\end{aligned}$$

当固定 $\boldsymbol{\alpha}$， $\boldsymbol{S}^{(t+1)}$ 去更新 $\boldsymbol{A}^{(t+1)}$ 和 $\boldsymbol{B}^{(t+1)}$ 时，可以得到：

$$\begin{aligned}&\left\|\boldsymbol{X} - \sum_{i=1}^{d}\alpha_i \boldsymbol{K}^{(i)}\boldsymbol{A}^{(t+1)}\boldsymbol{B}^{(t+1)}\right\|_F^2 + \lambda_1 \sum_{i,j}^{d}\left\|x_i^T\boldsymbol{A}^{(t+1)}\boldsymbol{B}^{(t+1)} - x_j^T\boldsymbol{A}^{(t+1)}\boldsymbol{B}^{(t+1)}\right\|_2^2 s_{i,j}^{(t+1)} + \lambda_2 \sum_{i,j}^{d}\left\|s_i^{(t+1)}\right\|_2^2 \\ &\le \left\|\boldsymbol{X} - \sum_{i=1}^{d}\alpha_i \boldsymbol{K}^{(i)}\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\right\|_F^2 + \lambda_1 \sum_{i,j}^{d}\left\|x_i^T\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)} - x_j^T\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\right\|_2^2 s_{i,j}^{(t+1)} + \lambda_2 \sum_{i,j}^{d}\left\|s_i^{(t+1)}\right\|_2^2\end{aligned}$$

由上面两式可得：

$$\begin{aligned}&\left\|\boldsymbol{X} - \sum_{i=1}^{d}\alpha_i \boldsymbol{K}^{(i)}\boldsymbol{A}^{(t+1)}\boldsymbol{B}^{(t+1)}\right\|_F^2 + \lambda_1 \sum_{i,j}^{d}\left\|x_i^T\boldsymbol{A}^{(t+1)}\boldsymbol{B}^{(t+1)} - x_j^T\boldsymbol{A}^{(t+1)}\boldsymbol{B}^{(t+1)}\right\|_2^2 s_{i,j}^{(t+1)} + \lambda_2 \sum_{i,j}^{d}\left\|s_i^{(t+1)}\right\|_2^2 \\ &\le \left\|\boldsymbol{X} - \sum_{i=1}^{d}\alpha_i \boldsymbol{K}^{(i)}\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\right\|_F^2 + \lambda_1 \sum_{i,j}^{d}\left\|x_i^T\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)} - x_j^T\boldsymbol{A}^{(t)}\boldsymbol{B}^{(t)}\right\|_2^2 s_{i,j}^{(t)} + \lambda_2 \sum_{i,j}^{d}\left\|s_i^{(t)}\right\|_2^2\end{aligned}$$

定理 1 设{ α_t }是由算法 2 产生的序列，那么对于 $\forall t \ge 1$，式(30)成立。

$$F(\boldsymbol{\alpha}_t) - F(\boldsymbol{\alpha}^*) \le \frac{2\gamma L\left\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}^*\right\|_F^2}{(t+1)^2} \tag{30}$$

根据参考文献[15]可知，γ 是事先定义的常量，L 是式(23)中 $f(\boldsymbol{\alpha})$ 梯度的 Lipschitz 常数以及 $\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha})$

通过上面的不等式和定理 1，能够很容易地看出本文的算法是收敛的。

3 实验结果和分析

3.1 实验数据集和对比算法

本文在六个数据集上测试 LS_NFS 算法的性能，分别为 Movements、Ecoli、Urban_land、Ionosphere、Colon、Lung_discrete。其中前三个均来自 UCI[16]，而后三个来自属性选择数据集[17]。数据集的详情如表 1 所示。

表 1 数据集信息统计

Table 1 Dataset information statistics

数据集	样本数	属性数	类数
Movements	360	90	15
Ecoli	336	343	8
Urban_land	168	147	9
Ionosphere	351	34	2
Colon	62	2000	2
Lung_discrete	73	325	7

本文的实验均在 Win7 系统下，MATLAB2014a 平台上运行测试。选择五种对比算法与本文提出的算法进行比较：LS[18]根据两个数据点很近，则它们很可能有很多相似的关系。通过计算其拉普拉斯分数来反映局部结构的保持能力。最后达到好的属性选择效果。CSFS[19]是一种凸半监督多标签属性选择算法，它主要用于大规模多媒体分析。它把不同属性之间的相互关系考虑在内，同时给无标签数据初始化一个为零的标签，最后最小化稀疏来进行属性选择。NetFS[20]是一个鲁棒的无监督属性选择算法，它将潜在的表示学习嵌入到属性选择中来减轻噪声的影响，从而达到好的效果。RUFS[21]也是一个鲁棒的无监督属性选择算法，它把聚类和属性选择同时进行，并且减少了算法的时间和空间复杂度。RSR[22]通过 $l_{2,1}$ − 范数来约束自我表示系数矩阵，从而选择具有代表的特征，并确保了对离群点的健壮性。

分析以上的算法，它们都各有各自的优点，但都没有考虑属性之间的非线性关系和相似性。而本文的算法通过结合局部结构学习与核方法来更充分地挖掘属性之间的关系，进而选择出更好的属性子集。

3.2 实验结果和分析

本文实验首先通过十折交叉验证来对数据集进行划分，用算法选好的属性构成新的属性集，然后再用 SVM 进行分类。最后用分类准确率来评估算法的性能。所有的算法都是在同一环境下进行，最后提取 10 次运行实验结果的均值加减均方误差来评价各算法的性能。

具体地，分类准确率定义如下：

$$acc = \boldsymbol{X}_{correct} / \boldsymbol{X} \tag{31}$$

其中：$\boldsymbol{X}$ 代表样本总数， $\boldsymbol{X}_{correct}$ 代表分类正确的样本数。同时定义标准差来衡量本文算法的稳定性，如下所示：

$$std = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(acc_i - \mu)^2} \tag{32}$$

其中：N 表示实验次数， acc_i 代表第 i 次实验的分类准确率，μ 代表平均分类准确率，std 越小，代表算法越稳定。

各个算法在六个数据集上实验结果对比如图 1 所示，具体数据结果如表 2 所示。

通过图 1 中的 6 个子图所示，能够很清楚地看到 LS_NFS 算法在 6 个数据集上的效果都是很好的。从 Ionosphere 和 Colon 两个数据集上，可以看出，对于二分类问题，LS_NFS 算法体现出了很好的性能。与此同时，在余下的四个多分类数据集上，LS_NFS 算法也有比较突出的表现。由于每个数

据集都有各自的特点，不同的算法可能适用不同的数据集。所以不能保证每次的结果都是最好的，但与其他的对比算法相比大部分情况下是较高的，最后的分类效果也是最好的。

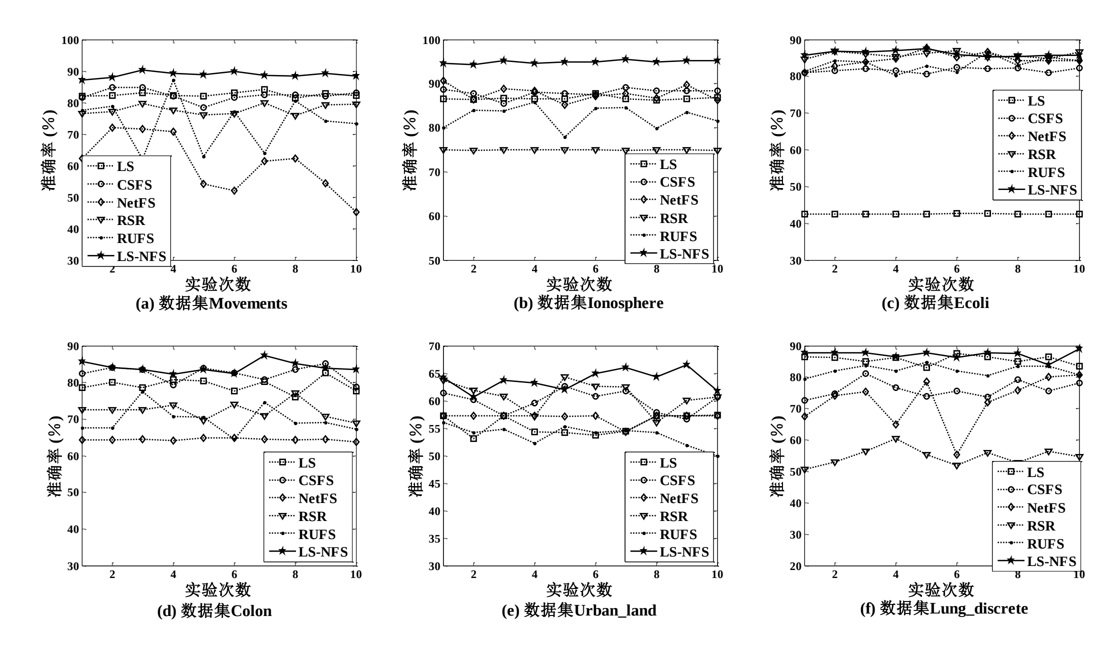


图 1　实验结果图

Fig. 1　Experimental results

表 2　准确率（均值±均方差）统计结果

Table 2　Accuracy (mean ± standard deviation) statistical results

数据集	LS	CSFS	NetFS	RUFS	RSR	**LS_NFS**
Movements	82.50±0.75	82.31±1.71	60.61±8.66	73.75±7.99	77.78±1.56	**88.81±0.86**
Ecoli	42.56±0.03	81.66±0.61	84.47±1.78	83.29±1.84	85.81±0.82	**86.19±0.69**
Urban_land	55.61±1.69	59.86±1.90	56.96±0.88	53.75±1.73	60.97±2.53	**63.77±1.79**
Ionosphere	86.69±0.39	87.92±0.93	87.70±1.61	82.52±2.44	74.94±0.02	**94.90±0.33**
Colon	79.24±1.83	82.40±2.02	64.38±0.29	69.79±3.56	72.29±2.31	**84.14±1.50**
Lung_discrete	85.59±1.34	76.11±2.55	72.45±7.48	82.20±1.57	54.73±2.67	**87.20±1.32**
平均	72.03±1.01	78.38±1.62	71.10±3.45	74.22±3.19	71.09±1.65	**84.17±1.08**

通过表 2 可以看出 LS_NFS 算法与其他五种对比算法的具体情况，与 LS 算法比较平均提高了 12.14%；比 CSFS 算法平均提高了 5.79%。此外，对比 NetFS、RUFS、RSR 算法，准确率分别提高了 13.07%、9.95%、13.08%。从标准差上来看，LS_NFS 算法在六个数据集上的平均标准差虽然不是最小的，但仅次于 LS 算法，只比 LS 算法的平均标准差高 0.07。这一定程度上也说明 LS_NFS 算法有较好的稳定性。LS_NFS 算法取得较好的效果，主要与以下两点有关：考虑了数据属性之间的相似性；充分的考虑了数据属性之间的非线性关系。

虽然不同类型的数据集的分布不同，且有一些干扰因素。但从实验结果来看，本文提出的 LS_NFS 具有最好的属性选择效果。相比其他对比算法，分类准确率也是最高的，同时也说明 LS_NFS 算法选出了最有代表性的属性。

4　结束语

本文通过考虑数据属性之间的相似性和非线性关系，提出了一种新的无监督非线性属性选择算法。即通过局部结构学习找出属性之间的相似性，然后通过核方法来找出数据属性之间的非线性关系，最后通过稀疏正则化因子来进行属性选择。在整个模型中还加入了低秩约束，来更好的完善提出的模型。比一般的属性选择算法更具有显著的挖掘效果。经实验结果证实，本文的算法在分类准确率和稳定性上都取得了很大的提升。在今后的工作中，本文尝试结合更前沿的理论改进算法。

参考文献：

[1] Zhu Xiaofeng, Suk H I, Shen Dinggang. Matrix-similarity based loss function and feature selection for Alzheimer's disease diagnosis [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2014: 3089-3096.

[2] Zhu Xiaofeng, Huang Zi, Shen Hengtao, *et al.* Dimensionality reduction by mixed kernel canonical correlation analysis [J]. Pattern Recognition, 2012, 45(8): 3003-3016.

[3] Zhu Xiaofeng, Zhang Shichao, Jin Zhi, *et al.* Missing value estimation for mixed attribute dataSets [J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(1): 110-121.

[4] Zhu Xiaofeng, Li Xuelong, Zhang Shichao. Block-row sparse multi-view multi-label learning for image classification. [J]. IEEE Trans on Cybernetics, 2016, 46(2): 450-461.

[5] Wei Xiaokai, Yu P S. unsupervised feature selection by preserving stochastic neighbors [C]// Proc of the 19th International Conference on Artificial Intelligence and Statistics 2016: 1.

[6] Zhu Xiaofeng, Huang Zi, Yang Yang, *et al.* Self-taught dimensionality reduction on the high-dimensional small sized data[J]. Pattern Recognition, 2013, 46(1): 215-229.

[7] Zhang Shichao, Jin Zhi, Zhu Xiaofeng. Missing data imputation by

utilizing information within incomplete instances[J]. Journal of Systems and Software, 2011, 84(3): 452-459.

[8] Nie Feiping, Zhu Wei, Li Xuelong. Unsupervised feature selection with structured graph optimization [C]//Proc of the 30th AAAI Conference on Artificial Intelligence. Palo Alto:AAAI Press, 2016: 1302-1308.

[9] Hou Chenping, Nie Feiping, Li Xuelong, *et al*. Joint embedding learning and sparse regression: a framework for unsupervised feature selection [J]. IEEE Trans on Cybern, 2017, 44(6): 793-804.

[10] Shao Weixiang, He Lifang, Lu C T, *et al*. Online unsupervised multi-view feature selection [C]//Proc of the 16th IEEE International Conference on Data Mining.Piscatraway,NJ: IEEE Press 2016: 1203-1208.

[11] Jian Ling, Li Jundong, Shu Kai, *et al*. Multi-label informed feature selection [C]//Proc of International Joint Conference on Artificial Intelligence. Palo Alto:AAAI Press, 2016: 1627-1633.

[12] Liu Xinwang, Wang Lei, Zhang Jian, *et al*. Global and local structure preservation for feature selection [J]. IEEE Trans on Neural Networks & Learning Systems, 2017, 25 (6): 1083-1095.

[13] Lu Canyi, Lin Zhouchen, Yan Shuicheng. Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization [J].IEEE Trans on Image Processing, 2015,24(2): 646-654.

[14] Wen Zaiwen, Yin Wotao. A feasible method for optimization with orthogonality constraints [J]. Mathematical Programming, 2013, 142 (1-2): 397-434.

[15] Nesterov Y. Introductory lectures on convex optimization [J]. Applied Optimization, 2004, 87(5): xviii, 236.

[16] UCI repository of machine learning datasets [EB/OL]. [2016-05-27]. http://archive. ics. uci. edu/ml/

[17] Feature selection datasets [EB/OL]. [2016-05-27]. http://feature selection. asu. edu/datasets. Php

[18] He Xiaofei, Cai Deng, Niyogi P. Laplacian score for feature selection. [C]//Proc of International Conference on Neural Information Processing Systems. Cambridge : MIT Press, 2005: 507-514.

[19] Chang Xiaojun, Nie Feiping, Yang Yi, *et al*. A convex formulation for semi-supervised multi-label feature selection [C]// Proc of the 28th AAAI Conference on Artificial Intelligence.Palo Alto: AAAI Press, 2014: 1171-1177.

[20] Li Jundong, Hu Xia, Wu Liang, *et al*. Robust unsupervised feature selection networked data [C]//Proc of SIAM International Conference on Data Mining. 2016: 387-395.

[21] Qian Mingjie, Zhai Chengxiang. Robust unsupervised feature selection [C]//Proc of International Joint Conference on Artificial Intelligence. 2013: 1621-1627.

[22] Zhu Pengfei, Zuo Wangmeng, Zhang Lei, *et al*. Unsupervised feature selection by regularized self-representation [J]. Pattern Recognition, 2015, 48(2): 438-446.